

PEARC22 BOF

July 12, 2022

Reproducibility and Trustworthiness of Scientific Research

H. Biral Runesha, Ana Trisovic, Sandra Gesing
and James Wilgenbusch



Our BoF today ...

1. Introduction and framing of the BOF
2. Panelists introductory comments
3. Questions and answers

BoF notes: <https://tinyurl.com/ynpz3yks>

Goals

This BoF will discuss opportunities and challenges for developing support services to expand the user base, lower barriers for capturing artifacts while doing research, and brainstorm how to work as a community towards a concerted effort to build an ecosystem of tools to support reproducibility.

BoF notes: <https://tinyurl.com/ynpz3yks>

Reproducibility and Replicability

- **Reproducibility:** Obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis
 - > not working on **numerical reproducibility**
- **Replicability:** obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data

What do we mean by reproducibility?

recreate

Re-compute

Replicate

regenerate

review

repeat

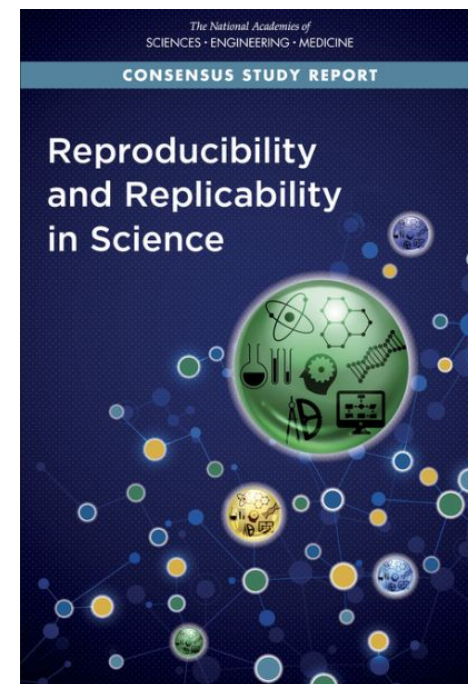
restore

recycle

revise

redo

rerun



Re-examine

repurpose

reconstruct

Improving Trustworthiness of Computational Results: Opportunities for the NSF Office of Advanced Cyberinfrastructure to address recommendations from the National Academies Report on Reproducibility

“OAC Reproducibility Opportunities Report”
Draft for Comment Summary

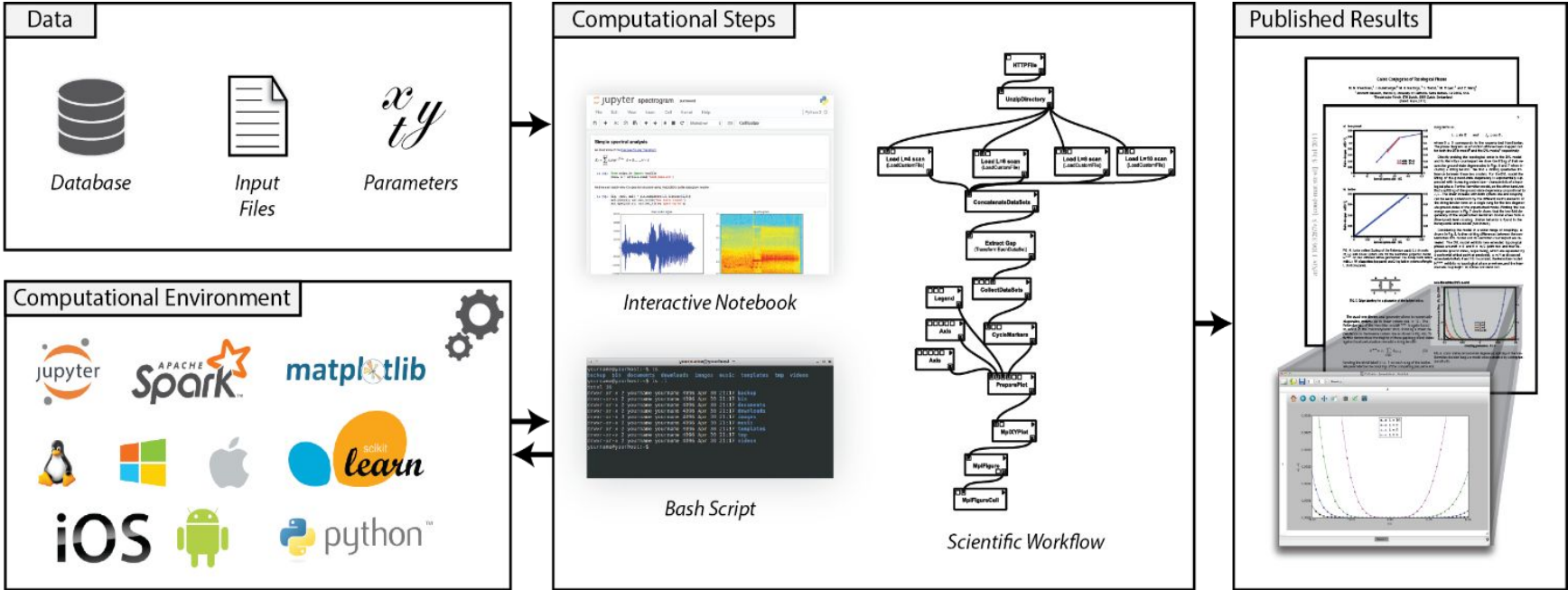
Wolfgang Bangerth, Juliana Freire, Patrick Heimbach, Michael Heroux (chair), Ivo Jimenez, Ellen Rathje, Hakizumwami Runesha, Victoria Stodden

Version for community comment:

https://docs.google.com/document/d/1d7kJ28-m8xxtrXQbTodKfFmDiR11uJto1jFb2h_w7bY/edit?usp=sharing

Vision for Trustworthy Computational Science

We look toward a future for computational science where all computational results are reproducible, including those from pipelines across multiple teams. Effective and efficient reproducibility will enable qualitative advances in science and make possible a new level of demonstrable trust in scientific results and outcomes.



SOURCE: Fernando Chirigati & Juliana Freire

What are the questions?


- Can you **reproduce the results** of a scientific research in a published paper?
- Do you **have enough information** to allow you to reproduce the results?
- Research **takes time** before getting results. Do we collect enough information **while doing research** to facilitate the reproducibility of the final results?
- Reproducibility is **hard and can be labor intensive**. How do we minimize the manual effort required to put together the artifact to be shared?
- etc.

Do existing tools and repositories fully address the reproducibility question? Are we using what is available today?

Examples of existing tools

- **Project jupyter:** web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
- **Containers:** software to package applications allowing them to be portable to any system running a Linux OS. It captures necessary system dependencies and vastly help with reproducibility
- **Github/Gitlab:** a web-based version-control and collaboration platform for software developers
- **Globus:** software for transferring and sharing files. It is also used to build applications and gateways
- **Digital Object identifier (DOI):** is a persistent identifier or handle used to identify digital objects uniquely.
- etc.

Example of initiatives

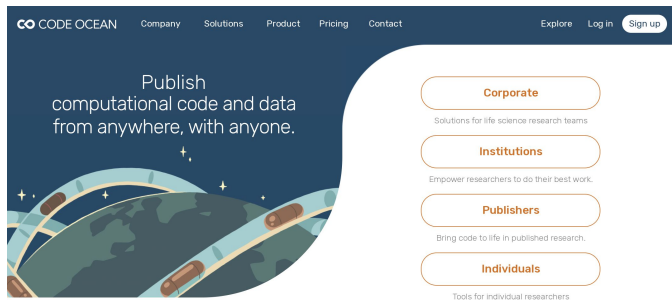


The Dataverse Project

Open source research data repository software

- Researchers**: Enjoy full control over your data. Receive web visibility, academic credit, and increased citation counts. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?
- Journals**: Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between articles in your journal and associated data. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?
- Institutions**: Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?
- Developers**: Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. Integrate research analysis, visualization and exploration tools, or other research and data archival systems with Dataverse. Want to contribute?

dataverse.org



CODE OCEAN

Publish computational code and data from anywhere, with anyone.

- Corporate**: Solutions for life science research teams
- Institutions**: Empower researchers to do their best work.
- Publishers**: Bring code to life in published research.
- Individuals**: Tools for individual researchers

Code Ocean transforms computational research.

We've created an integrated computational research platform for increased productivity, guaranteed computational reproducibility, and seamless collaboration. The platform captures the exact version of code, data, and development environment that generated every result. It also

codeocean.com



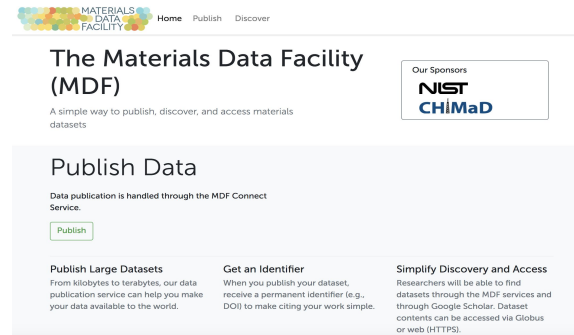
WHOLE TALE

Reproducibility Simplified

Use Whole Tale to empower and share your research

- Use Whole Tale to create and publish your own transparent and reproducible research.
- Explore existing reproducible research created using Whole Tale.
- Learn more about Whole Tale, an open source platform for reproducible research.

wholetale.org



MATERIALS DATA FACILITY

Home Publish Discover

The Materials Data Facility (MDF)

A simple way to publish, discover, and access materials datasets

Our Sponsors

NIST
CHIMaD

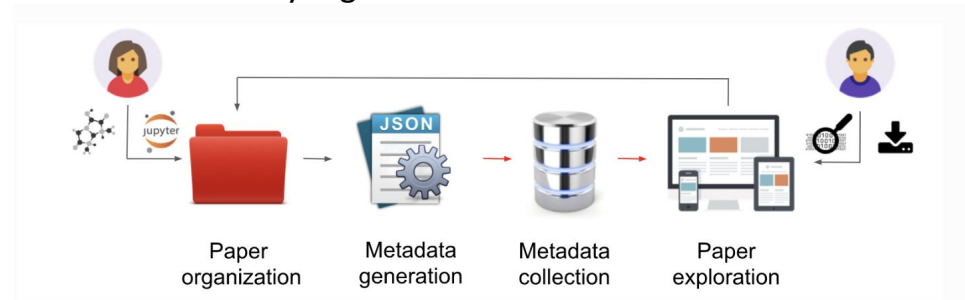
Publish Data

Data publication is handled through the MDF Connect Service.

Publish

- Publish Large Datasets**: From kilobytes to terabytes, our data publication service can help you make your data available to the world.
- Get an Identifier**: When you publish your dataset, receive a permanent identifier (e.g., DOI) to make citing your work simple.
- Simplify Discovery and Access**: Researchers will be able to find datasets through the MDF services and through Google Scholar. Dataset contents can be accessed via Globus or web (HTTPS).

materialsdatafacility.org



etc.

qresp.org

Observations

- Some existing tools are addressing **parts of the reproducibility question**
- some are focusing on reproducible papers, i.e. allowing **interaction with final results** of research, rather than **an entire workflow**
- some projects focus on **project management** rather than computational reproducibility
- Some require users to **manually** define workflows of their research (not automatically)
- eb-based tools that support the use of containers
- the artifact needed for reproducibility is put together **after the research is completed**
- Some are enhanced **repositories.**

RCC Data Hub

<https://datahub.rcc.uchicago.edu/>

A **data portal** to search, view and download workflows, tools, documentation, and all data sets needed to **reproduce the results of a scholarly work**.

Aims:

- **Advance the openness of all scientific data** produced throughout the life cycle of a project for compliance with funded research grants and accelerated productivity.
- **Increase the integrity and reproducibility of scientific results.**

The screenshot displays the RCC Data Hub website interface. At the top, there is a dark red navigation bar with the University of Chicago logo and 'DataHub' text, along with links for HOME, SEARCH, DOCUMENTATION, and CONTACT US. Below this, the main content area is divided into several sections:

- Publication Criteria:** A sidebar on the left contains search filters for Principal Investigator (All), DOI (Enter Paper DOI), Paper Title (Enter Paper Title), and Publication Name (All). Search and Clear buttons are at the bottom.
- The RCC Data Hub:** A central section with a header image showing a 'DATA' sign and descriptive text: 'The Research Computing Center experimental. This includes data documenting, storing, and sharing results of researchers' publications. The RCC Data Hub uses RCC's in...'.
- DMREF | Datahub:** A section with a dark header and navigation links (ABOUT, SEARCH, CONTACT US). It features a search bar and a 'Show 10 entries' dropdown.
- Publication List:** A table of search results with columns for Paper Title, Paper Author(s), Published in, and Published Date. Each entry includes buttons for 'Figures/ Tables', 'Notebook', 'PDF Source', and 'Download'.

Paper Title	Paper Author(s)	Published in	Published Date
Age and structure of a model vapour-deposited glass	Daniel R. Reid, Ivan Lyubimov, M. D. Ediger, Juan J. de Pablo	Nature Communications	2016-10-20
Effect of Low-Concentration Polymers on Crystal Growth in Molecular Glasses: A Controlling Role for Polymer Segmental Mobility Relative to Host Dynamics.	C. Huang, C. T. Powell, Y. Sun, T. Cai, Lian Yu	Journal of Physical Chemistry B	2017-01-31
Highly organized smectic-like packings in vapor-deposited glasses of a liquid crystal	Ankit Gujral, Jaritza Gómez, Jing Jiang, Chengbin Huang, Kathryn A. O'Hara, Michael F. Toney, Michael L. Chabinyk, Lian Yu, and M. D. Ediger	Chemistry of Materials	2016-12-26

What are we trying to achieve?

As a community, can we work together to build an ecosystem of tools and service to help researchers capture and collect metadata/info about the methodology, data, software, tools, platform, etc. , associated with results, while doing their research with the least amount of effort?

Rethinking how to support Research Computing and training

We need to ...

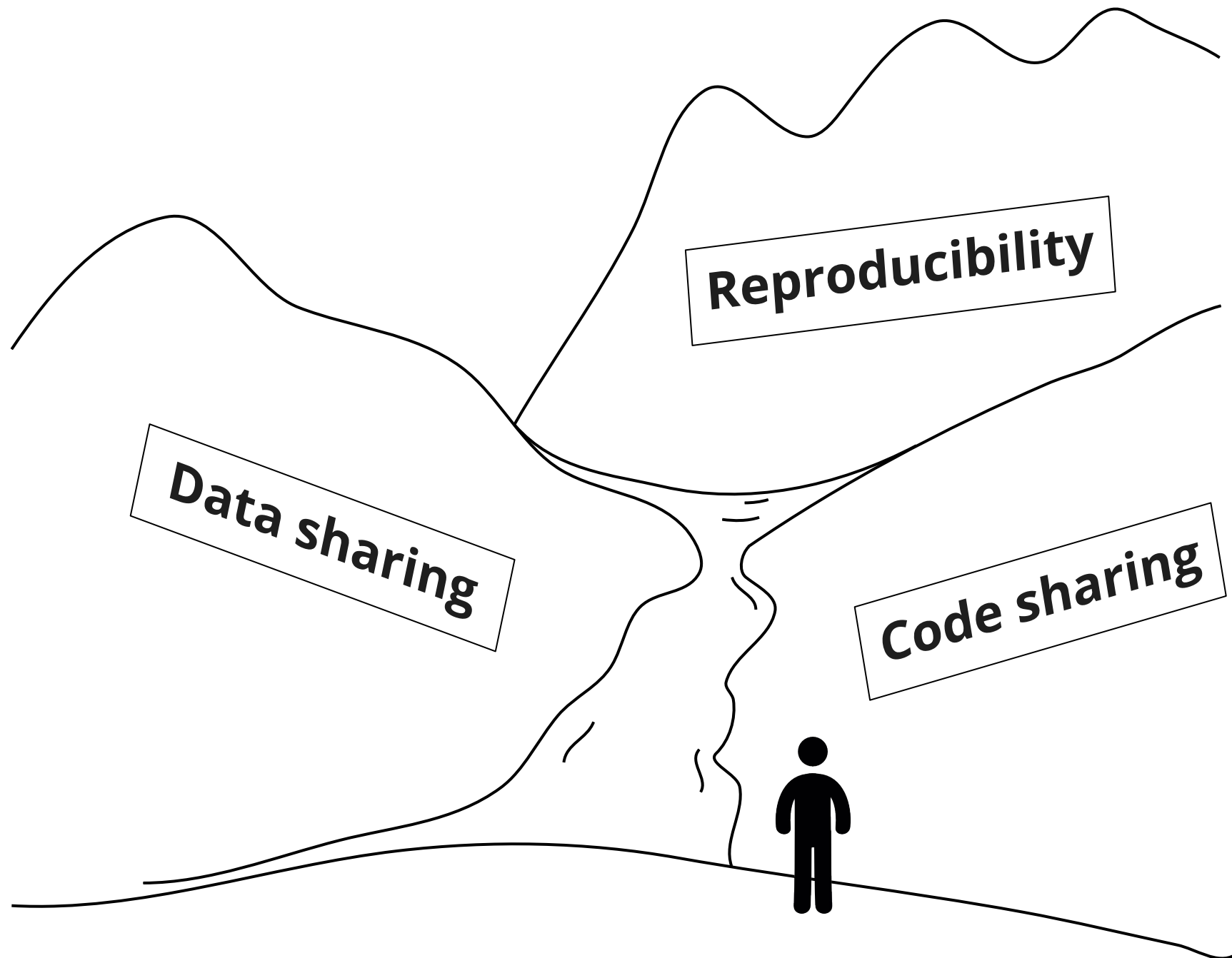
- **platforms and tools that facilitate reproducibility**
- **change how we currently conduct research**
- **develop new training and services**

Our BOF today ...

1. Introduction and framing of the BOF
2. Panelists introductory comments
3. Questions and answers

Prerequisites: FAIR + R

- Reproducibility requires some level of sharing (doesn't need to be open access)
- FAIR principles - guidelines for sharing data and, more recently, research software
 - Data repositories and software repositories implement features that meet FAIR principles
 - Repositories aim to facilitate computational reproducibility



Research repositories

zenodo

Dataverse[®]



DRYAD



GitHub



figshare



GitLab

Provenance capture

ReproZip



WHOLE TALE



Encapsulation tools

colab



CODE OCEAN

binder



RENKU

Stencila



Workflow engines



nextflow



Apache Airflow



COMMON WORKFLOW LANGUAGE

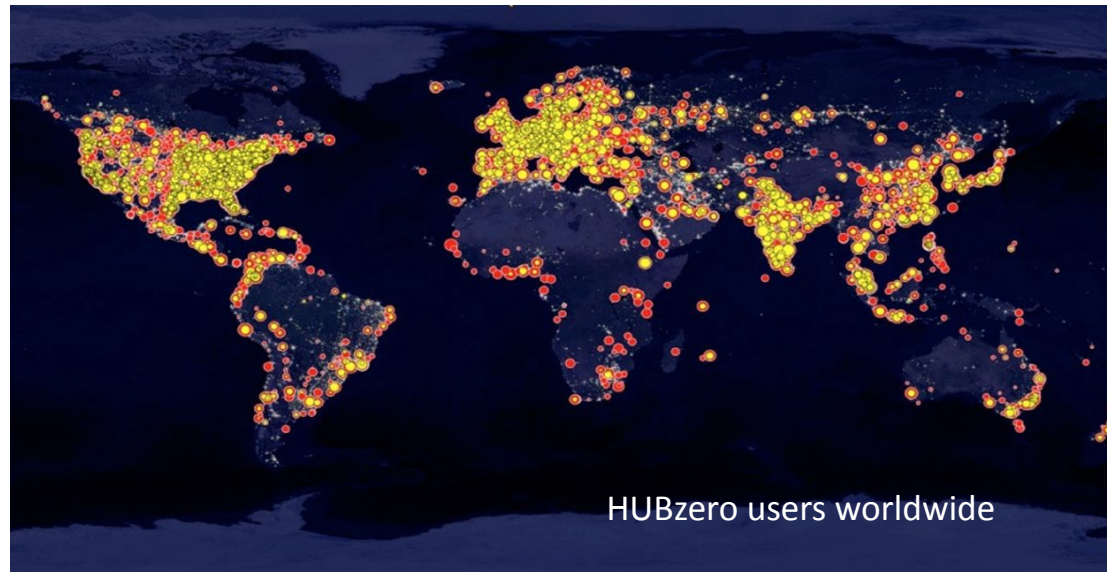
Science Gateway Technologies

- Widely used complete frameworks (HUBzero, Open Science Framework, Galaxy, Globus, Data Portal, etc.)
- Based on different stacks of products / Open standards / languages in back-end (Perl, Python, etc.) with provision of hardware in the

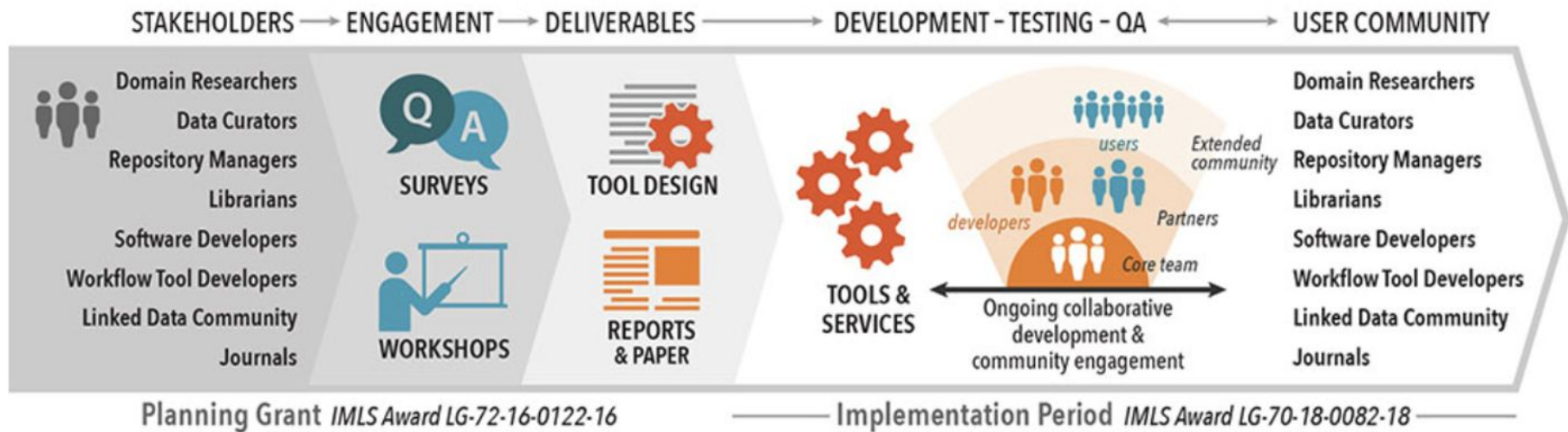
Predestined for sharing data and computational methods and reproducibility

in ONE instance

Sharing between different instances and technologies is complex



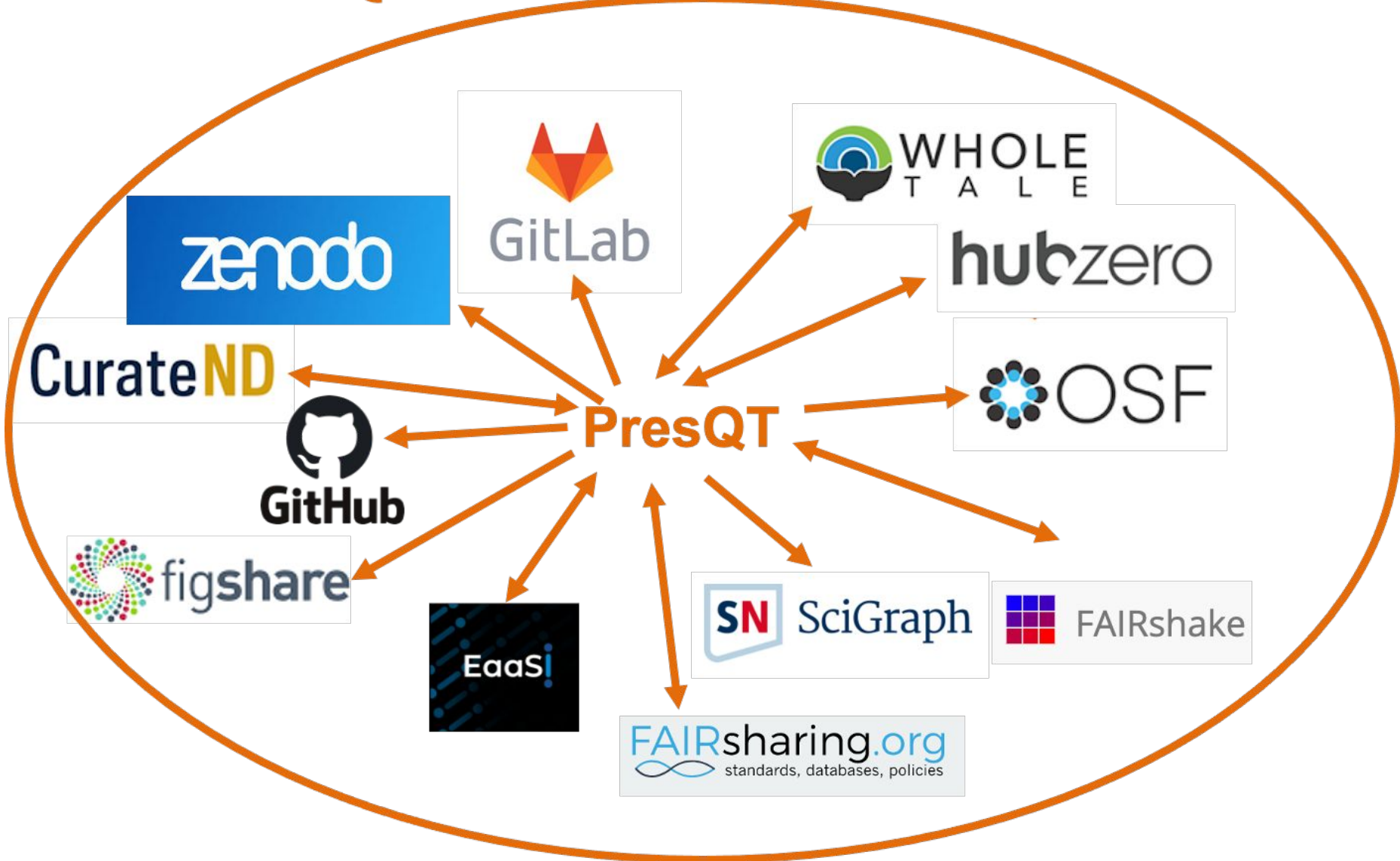
PresQT Connects and Enhances



Concept

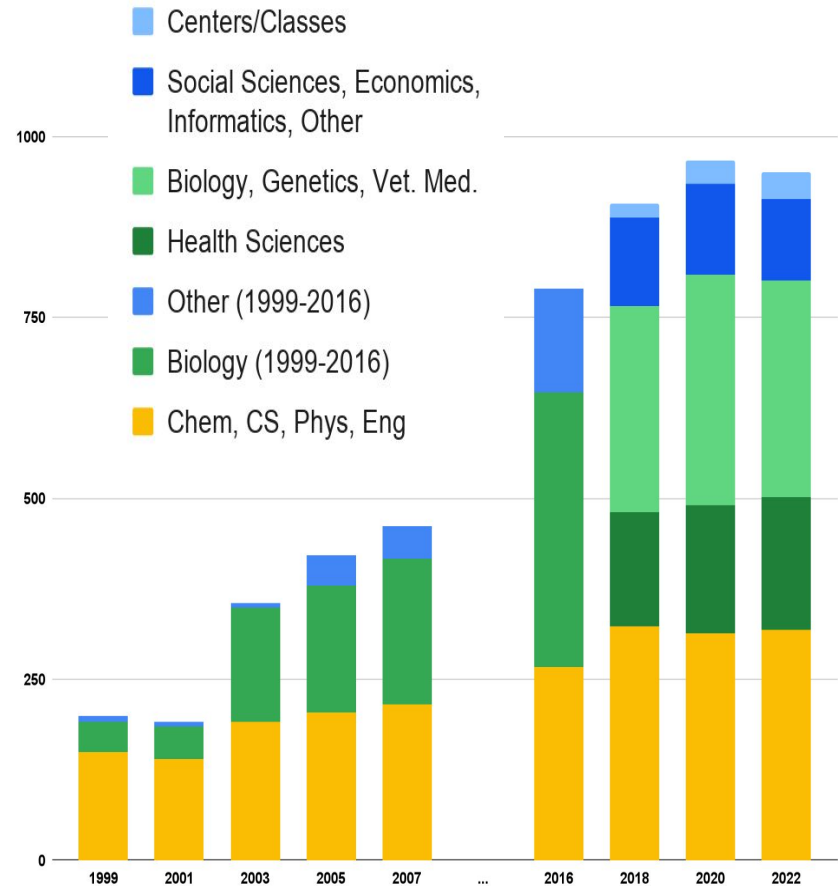
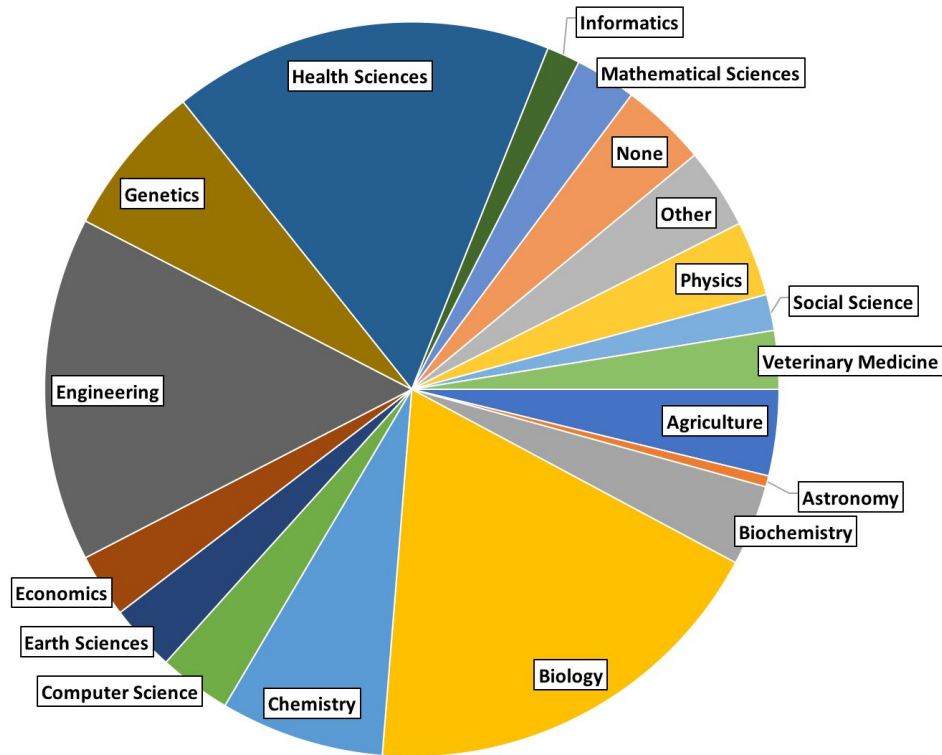
- **not** standalone solutions
- partner systems and services easily integrable via RESTful APIs and services
- user-centered open design and collaborative development

PresQT Connects and Enhances



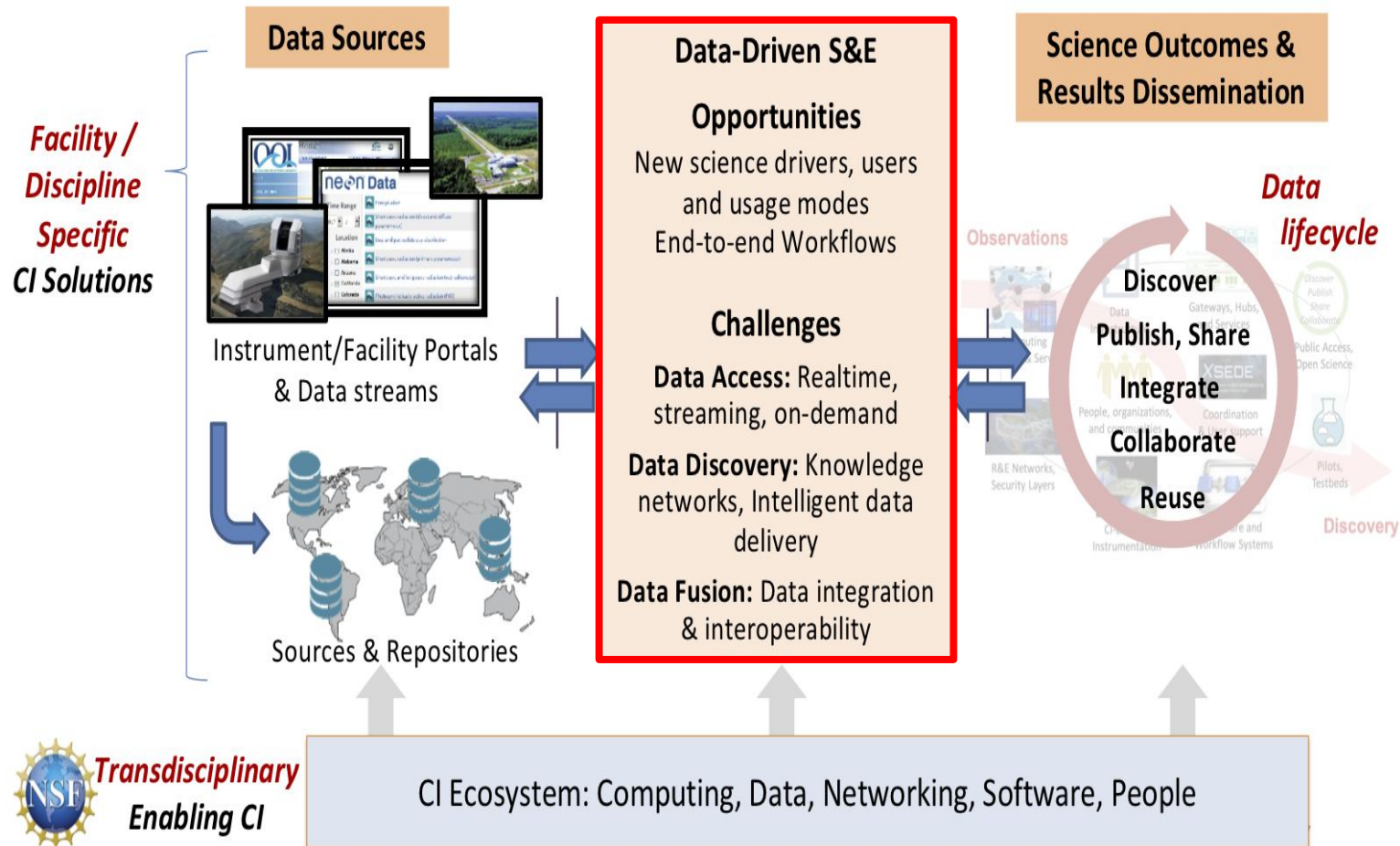
The Long Tail of Research Computing

Groups in 2020: 900 User Groups
4,555 Active users



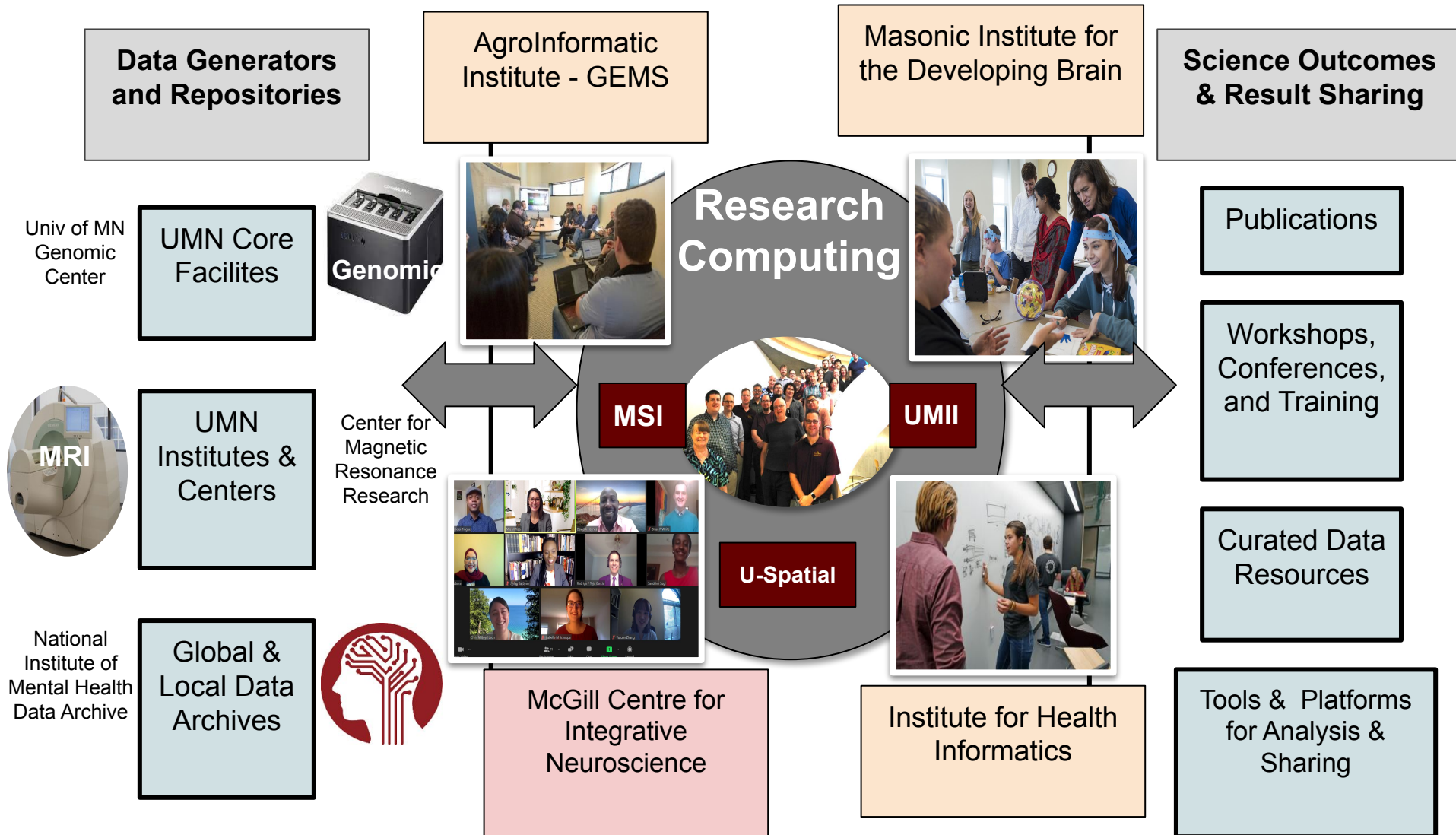
Biggest increasing in Life Sciences

Data-Intensive Discovery Pathways – The “missing middle”



*Slide Credit: Manish Parashar, Office Director of the [Office of Advanced Cyberinfrastructure \(OAC\)](#) at the [National Science Foundation \(NSF\)](#) presented at the Fall Midwest Big Data Hub All hands Meeting, October 30, 2019.

Developing the “Missing Middle” System-wide View of the Research Computing

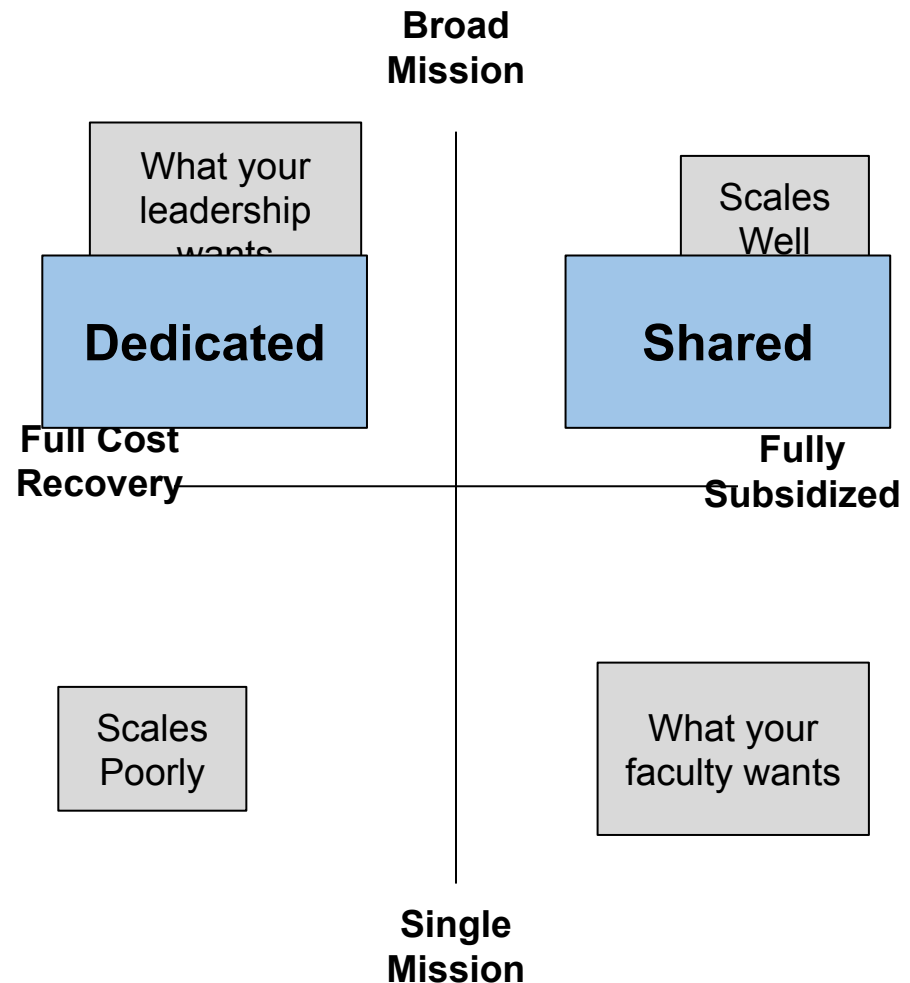


Key Ingredients to Reproducibility

Things

Scalable, multipurpose systems infrastructure

- **Shared** - Available to everyone on a first come first serve basis
- **Dedicated** - Dedicated to single research group or project

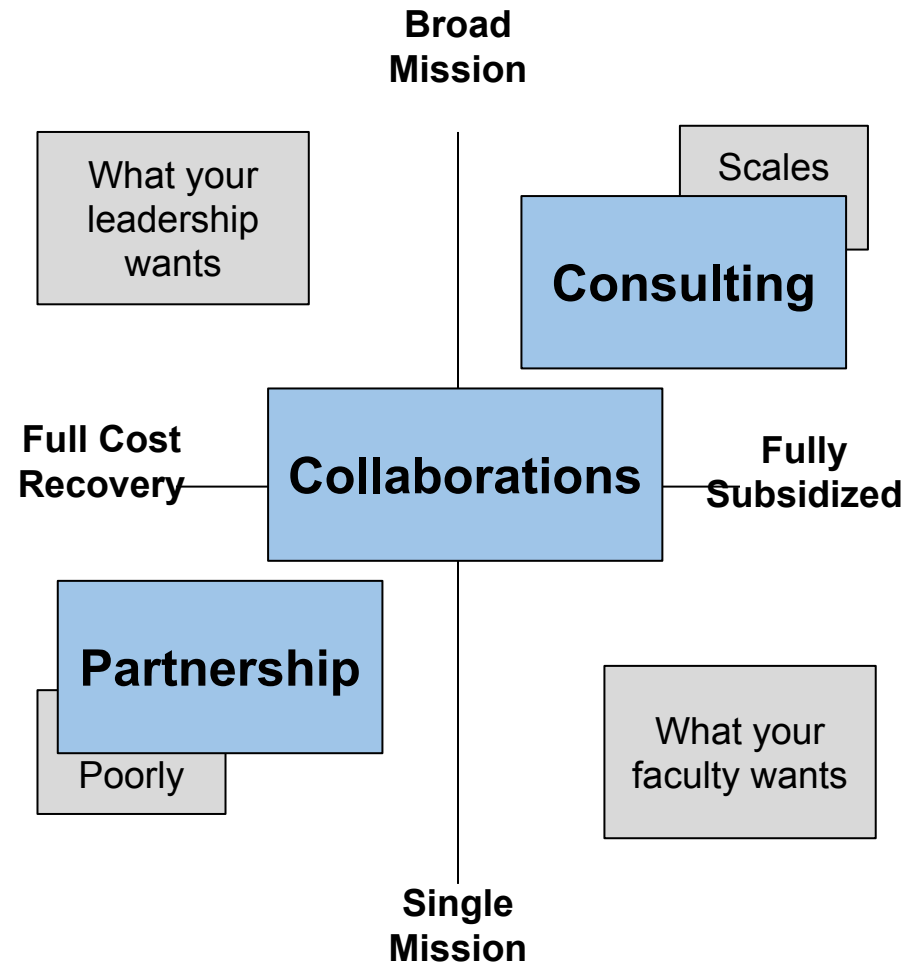


Key Ingredients to Reproducibility

People

Scalable, multipurpose staff experts

- **Consultations** - No direct cost. Short engagements (<1 hour)
- **Collaborations** - Shared investments in emerging or locally specialized areas of research. Weeks to months of investment
- **Partnerships** - Cost recovery, well defined agreements, and longer term commitments for funding areas of research that are generally considered to be core UMN strengths.



Our BOF today ...

1. Introduction and framing of the BOF
2. Panelists introductory comments
3. Discussion – Questions and answers